УДК 519.854.2

ЗАСТУСУВАННЯ ПРИНЦИПІВ БАГАТОВИМІРНОЇ ПОЛІНОМИАЛЬНОЇ РЕГРЕСІЇ ДЛЯ РОЗВІДУВАЛЬНОГО АНАЛІЗУ ДАННИХ ТА ЗНАХОДЖЕННЯ ЛІНІЇ РЕГРЕСІЇ

Коваленко Д. А.

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Україна, Київ

Розглядається задача розвідувального аналізу даних та знаходження лінії регресії. Дана задача є універсальною и для її розв’язку використовується численні методи та математичні апарати. У даній роботі розглядається застосування принципів багатовимірної поліноміальної регресії для ознайомлення з даними та побудови регресійної моделі. Наведено змістовну та математичну постановки задачі, що розглядається. Запропоновано алгоритм застосування принципів багатовимірної регресії для розвідувального аналізу та знаходження лінії регресії. Наведені ключові аспекти реалізації та порівняльна таблиця алгоритмів.

Ключові слова: регресія, багатовимірна поліноміальна регресія, розвідувальний аналіз даних, кореляційна матриця.

магистрант, Коваленко Д. А. Применение принципов многомерной полиномиальной регрессии для разведывательного анализа данных и нахождения линии регрессии / Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского», Украина, Киев

Рассматривается задача разведывательного анализа данных и нахождения линии регрессии. Данная задача является универсальной и для ее решению используется многочисленные

методы и математические аппараты. В данной работе рассматривается применение принципов многомерной полиномиальной регрессии для ознакомления с данными и построения регрессионной модели. Приведены содержательную и математическую постановки задачи, рассматривается. Предложен алгоритм применения принципов многомерной регрессии для разведывательного анализа и нахождения линии регрессии. Приведенные ключевые аспекты реализации и сравнительная таблица алгоритмов.

Ключевые слова: регрессия, многомерная полиномиальная регрессия, разведывательный анализ данных, корреляционная матрица.

undergraduate, Kovalenko D applying multiple polynomial regression principles for exploratory data analysis and regression analysis / National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, Kyiv

The problem of exploratory data analysis and finding the regression line is considered. This problem is universal and many methods and mathematical approaches are used to solve it. In this paper, we consider the application of the principles of multidimensional polynomial regression to get acquainted with the data and construct a regression model. The content and mathematical formulation of the problem are provided. An algorithm for applying multidimensional regression principles for exploratory analysis and finding the regression line is proposed. The resulted key aspects of realization and the comparative table of algorithms

Key words: regression, multidimensional polynomial regression, reconnaissance data analysis, correlation matrix.

Вступ. Проблема знаходження істинної закономірності за результатами експериментів є універсальною. Немає ні однієї області діяльності людини, в якій так чи інакше не виникала б ця задача. В економічних, соціологічних та природничих науках часто вирішують задачу виявлення чинників, що визначають рівень і динаміку процесів. Таке завдання найчастіше вирішується методами кореляційного, регресійного, факторного і компонентного аналізу. Завдання регресійного аналізу полягає в побудові моделі, що дозволяє за значеннями незалежних показників отримувати оцінки значень залежної змінної. Різні аспекти розв'язку цієї проблеми розглядаються в таких науках, як математична статистика, теорія управління, теорія штучного інтелекту. В рамках теорії імовірності ця задача формулюється як оцінка лінії регресії по результатам статистичних експериментів і на практиці є областю прикладного регресійного аналізу.

Проблема відтворення невідомої залежності формулюється як класична задача прикладного регресійного аналізу: відтворення багатовимірної поліноміальної регресії по надлишковому опису і з довільно розподіленою похибкою. По результатам активних експериментів необхідно знайти невідомі коефіцієнти, частина з яких тотожно дорівнює нулю і невідома досліднику. На відміну від кореляційного аналізу не з'ясовує чи істотний зв'язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії. Регресійний аналіз використовується в тому випадку, якщо відношення між змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що може приймати цільова (залежна) змінна, яка обчислюється на заданому наборі значень вхідних (незалежних) змінних. У найпростішому випадку для цього використовуються

стандартні статистичні методи, такі як лінійна регресія. На жаль, більшість реальних моделей не вкладаються в рамки лінійної регресії. Наприклад, розміри продажів чи фондові ціни дуже складні для передбачення, оскільки можуть залежати від комплексу взаємозв'язків множин змінних. Таким чином, необхідні комплексні методи для передбачення майбутніх значень.

Саме тому розробка алгоритмів, які б допомогли вирішити цю проблему - проблему регресії багатьох змінних - є дуже актуальною у наш час і залишатиметься такою ще довго. Задача, яка постає перед нами, є дуже складною, адже загальний опис ситуації, який було зазначено вище, не показує усіх можливих складностей відтворення реальної залежностей складних процесів. Саме тому і досі не існує алгоритму, який міг би знайти дуже гарний розв’язок для усіх формулювань цієї задачі за прийнятний час. Проте вже було проведено дослідження та розроблено деякі можливі алгоритми розв’язання даної задачі, які для деяких варіації дають дуже гарні результати. Необхідно продовжувати дослідження та намагатися покращувати вже отримані результати.

Аналіз останніх досліджень і публікацій. Наразі розв’язування задач знаходження істинної закономірності за результатами експериментів є популярною і розв’язується багатьма напрямками комп’ютерних наук, в яких ведуться активні дослідження.

Найбільш популярній напрям розвитку – це напрям *машинного навчання*, що зводиться до знаходження оптимального розв’язку ітеративними евристичними алгоритмами. В принципі цю задачу може виконати будь-який алгоритм навчання з учителем. Наприклад, в [1], [2] надається приклад використання персептрона для вирішення задачі регресії. Хоча багато-шаровий персептрон і вирішую задачу

класифікації, його все ж можна використовувати для задач регресії не використовуючи при цьому активуючу функцію на останньому шарі – вихідні значення при цьому неперервні.

Дерева прийняття рішень можуть бути використані для задачі регресії. При цьому листки дерева приймають неперервні значення. В такому випадку такі дерева називають регресійними [3], [4].

Метод групового урахування аргументів (МГУА) розробляється академіком НАНУ О.Г. Івахненком та його школою, починаючи з 60-х років минулого століття [5], [6], [7]. Це типовий метод індуктивного моделювання і один з найбільш ефективних методів структурно-параметричної ідентифікації складних об'єктів, процесів і систем за даними спостережень в умовах неповноти інформації.

Метод адаптивних регресійних сплайнів (MARS) це форма регресійного аналізу що була запропонована Жеромом Фрідманом в 1991р. [8], [9]. Це непараметризований метод регресії і може розглядатися як розширення лінійних моделей що автоматично моделює нелінійні відношення та взаємодію між змінними. При цьому MARS розбиває вибірку на під-множити і використовує методу лінійного регресійного аналізу.

Методи LOESS та LOWESS (locally weighted scatterplot smoothing) [8], [9] – це два схожих непараметричних методи регресії які поєднують у собі декілька регресійних моделей у мета-моделі К-найближчих сусідів. LOESS це узагальнений метод LOWESS. Обидва методи будуються на класичних методах, такі як лінійний та нелінійний метод найменших квадратів. Вони адресують проблему коли класичні методи не можуть адекватно оцінити дані. LOESS об’єдную простоту лінійного методу найменших квадратів та гнучкість нелінійної регресії. Це стає можливим завдяки застосуванню простих

регресійних моделей на підмножинах вибірки для побудови функції що пояснює детерміновану частину даних точка за точкою. Краса цього методу полягає у тому що аналітику не потрібно задавати глобальну функцію якої-небудь форми – лише задавати прості моделі для сегментів даних.

Заслуговують на увагу і *методи регуляризації* [10], [11], [12]. Регулярізація, в математиці і статистиці, а також в задачах машинного навчання і обернених задачах, означає додавання деякої додаткової інформації, щоб знайти рішення некоректно сформованої задачі, або щоб уникнути перенавчання.

Дана робота базується на алгоритмі багатовимірної поліноміальної регресії при активному експерименті. В [13], [14], [15] описується алгоритм знаходження коефіцієнтів поліномів які є частково відомі досліднику. В даній роботі використовується частина знайдених результатів та застосовується разом з одновимірним регресійним аналізом.

З наведеного вище можна зробити висновок, що розробка та аналіз алгоритмів для знаходження істинної залежності по експериментальним даним має практичну цінність, оскільки задача є універсальною.

Мета та завдання статті. Метою даної роботи є проведення розвідувального аналізу з використанням принципів багатовимірного поліноміального аналізу. Для досягнення даної мети необхідно виконати наступні завдання:

– навести опис використаних даних;

– провести конвенційний розвідувальний аналіз;

– провести розвідувальний аналіз з багатовимірними поліномами;

– провести одновимірний регресійний аналіз та регресійний

аналіз з багатовимірними поліномами, порівняти результаті регресійного аналізу;

Опис даних. The World Happiness Report є опитуванням стану глобального щастя. Перший звіт був опублікований у 2012 році, другий - у 2013 році, третій у 2015 році, а четвертий - в оновленні 2016 року. Звіт 2017 року, який нараховує 155 країн за рівнем щастя, був випущений ООН під час святкування Міжнародного дня щастя 20 березня. Звіт продовжує отримувати глобальне визнання, оскільки уряди, організації та громадянське суспільство все частіше використовують показники щастя для інформування своїх рішень про політику. Провідні експерти з усіх галузей - економіки, психології, соціології, статистики, здоров'я, державної політики тощо - описують, як можна ефективно вимірювати добробут для оцінки прогресу націй. Звіти розглядають стан щастя в сучасному світі та показують, як нова наука про щастя пояснює особисті та національні варіації щастя.

Оцінки щастя та рейтинги використовують дані зі світового опитування Gallup. Оцінки базуються на відповідях на основне питання оцінки життя, задане в опитуванні. Бали складаються з національно-репрезентативних зразків на 2013-2016 роки та використовують ваги Галлапа, щоб зробити прогноз репрезентативним. Колонки, що слідують за оцінкою щастя, оцінюють ступінь, в якій кожна з шести чинників - економічне виробництво, соціальна підтримка, тривалість життя, свобода, відсутність корупції та щедрість - сприяють підвищенню якості життя в кожній країні.

Розглянемо змінні що будуть використовуватися в подальшому аналізі:

Таблиця 1

Опис змінних у наборі даних

Назва змінної	Опис змінної	Змінна
Country	Назва країни	
Happiness Score	Суб'єктивна метрика щастя зібрана методом опитування	y
GDP per Capita	Дохід на душу населення	x_1
Family	Середня кількість дітей в сім'ї	x_2
Life Expectancy	Очікувана тривалість життя	x_3

Розвідувальний аналіз

Розглянемо базові статистики незалежної змінної Happiness Score

Таблиця 2

Статистики незалежної змінної

Кількість спостережень	155
Середнє арифметичне	5.35
Стандартне відхилення	1.13
Мінімальне значення	2.69
Перший квантиль	4.5
Медіана	5.27
Третій квантиль	6.1
Максимальне значення	7.53
Асиметрія	0.0095
Коефіцієнт ексцесу	-0.75

Розглянемо розподіл незалежної змінної:

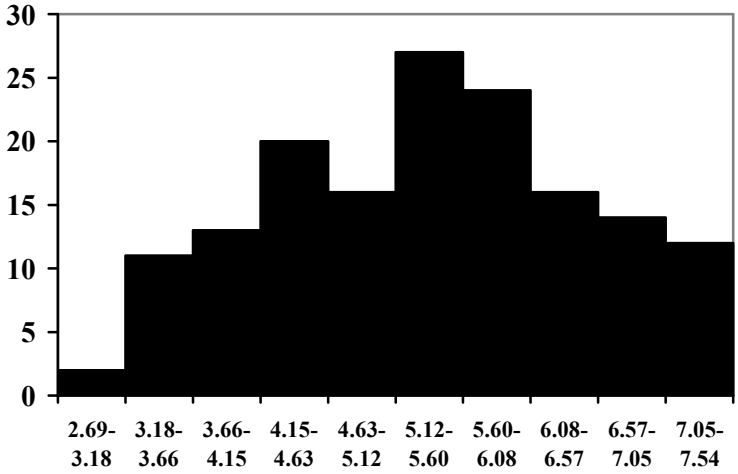

Рис. 1 - Розподіл щастя

Схоже на те, що змінна розподілена нормально, без асиметрії та від'ємним

коефіцієнтом ексцесу.

Наступним кроком поглянемо на залежність змінної Y від незалежних змінних:

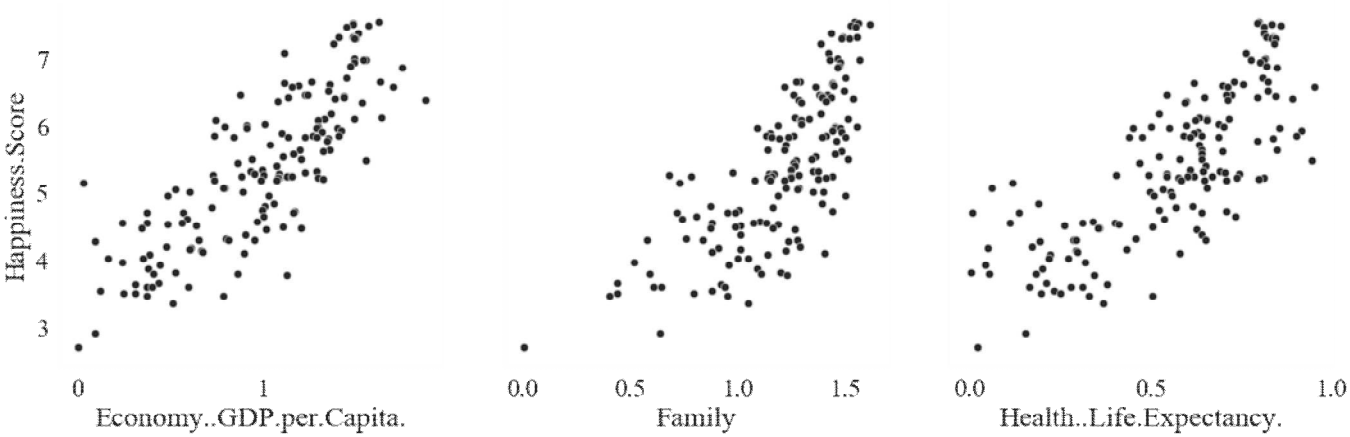

Рис. 2 - Точкова діаграма залежності коефіцієнту щастя від незалежних змінних

Після попереднього аналізу: ми можемо сказати, що:

- Дохід на душу населення та кількість дітей мають лінійний вплив на світове щастя;

- Очікувана тривалість життя, можливо, має квадратичний (експоненційний) вплив на світове щастя;

Розвідувальний аналіз з багатовимірними поліномами

Процеси у природі та соціумі можуть бути набагато складнішими ніж може здаватися на перший погляд. У таких випадках лінійні методи аналізу та моделі недостатньо точно описують реальні процеси. Саме тому є сенс використовувати квадратичні моделі та поліноми багатьох змінних. У даному розділі ми розглянемо більш складні залежності, як от залежність $Happiness.Score$ від $Family \times LifeExpectancy$.

При лінійному регресійному аналізі, лінія регресії має вигляд:

$$y = c_1 x_1 + c_2 x_2 + c_3 x_3$$

Припустимо, що залежна змінна може квадратично залежати від незалежних змінних або їх комбінацій. У такому випадку лінія регресії має вигляд:

$$y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_1 x_2 + c_5 x_1 x_3 + c_6 x_2 x_3 + c_7 x_1^2 + c_8 x_2^2 + c_9 x_3^2$$

Розглянемо графік залежності y від комбінованої змінної $x_2 x_3$:

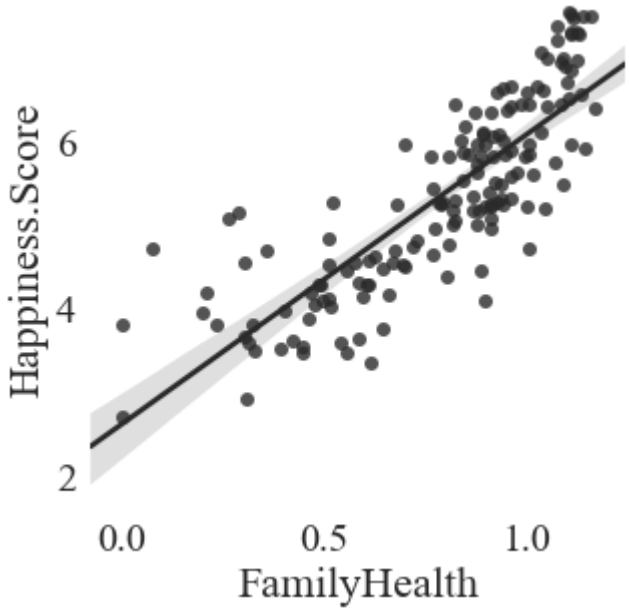

Рис. 3 – Залежніть світового щастя від комбінації розміру сім'ї та тривалості життя

Регресійний аналіз

Проведемо одновимірний регресійний аналіз та порівняємо результати за допомогою метрики R^2 обрахованої на тестовій виборці:

Таблиця 3

Результати роботи регресійних аналізів

Метод регресії	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	R^2
Початкові змінні	0.46	0.59	0.72	-	-	-	-	-	-	0.58
Поліноми	2.91	3.46	-1.58	0.85	1.52	-0.13	-0.4	-0.16	0	0.53

З даних таблиці 3 можна зробити висновок, що для даного набору даних додавання комплексних змінних покращує регресійну модель.

Висновки. Наведено постановку задачі розвідувального аналізу та регресії. Проведено розвідувальний аналіз, запропоновано алгоритм регресії що базується на принципах багатовимірної. Наведено результати регресійного аналізу. Запропонований алгоритм регресійного аналізу показав кращі результати.

Література:

1. Grendander U., Rosenblatt M., Statistical analysis of stationary time series, New York, 1957.

2. "Learning representations by back-propagating errors." Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams.

3. James H. Stock, Mark W. Watson. Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals // Introduction to

Econometrics. — 3. — Addison-Wesley, 2011. — P. 163-164. — 785 p. — ISBN 0138009007.

4. Norman Richard Draper, Harry Smith. “Applied Regression Analysis” Wiley, 1998.

5. Ивахненко А. Г. Метод группового урахування аргументів - конкурент методу стохастичної апроксимації // Автоматика. - 1968. - № 3. - С. 58-72.

6. Ивахненко А. Г. Системы эвристической самоорганизации в технической кибернетике. - Киев: "Техніка", 1971. - 392 с.

7. Ивахненко А. Г. Долгосрочное прогнозирование и управление сложными системами. - Киев: "Техніка", 1975. - 311 с.

8. Multivariate adaptive regression splines - Wikipedia [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines

9. Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines". The Annals of Statistics. 19: 1.

10. L1 and L2 Regulatization methods [Електронний ресурс] – Режим доступу до ресурсу: https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

11. Regularization (Mathematics) - Wikipedia [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Regularization_(mathematics)

12. Avoiding overfitting with regulatization [Електронний ресурс] – Режим доступу до ресурсу: https://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/

13. МЗ Згуровский, АА Павлов, "Принятие решений в сетевых системах с ограниченными ресурсами: Монография", К.: Наукова думка,–2010.–573 с

14. Згуровский М. З., Павлов А. А., Мисюра Е. Б., Мельников О. В. Методы оперативного планирования и принятия решений в сложных организаци- онно-технологических системах // Вісник НТУУ "КПІ". Інформатика, управління та обчислювальна техніка. К.: "ВЕК+", 2010.– No50

15. Павлов А. А., Калашник В. В., Коваленко Д. А. Построение багатовимірної полиномиальной регресії. Регресія при данных з повторяющимися аргументами // Вісник НТУУ "КПІ". Серія «Інформатика, управління та обчислювальна техніка». – К.: "ВЕК+", 2015. – No63. – 4 с.

References:

1.Grendander U., Rosenblatt M., Statistical analysis of stationary time series, New York, 1957.

2. "Learning representations by back-propagating errors." Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams.

3. James H. Stock, Mark W. Watson. Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals // Introduction to Econometrics. — 3. — Addison-Wesley, 2011. — P. 163-164. — 785 p. — ISBN 0138009007.

4. Norman Richard Draper, Harry Smith. "Applied Regression Analysis" Wiley, 1998.

5. Ivakhnenko A. G. Metod grupovogo urakhuvannya argumentiv - konkurent metodu stokhastichnoї aproksimatsiї // Avtomatika. - 1968. - № 3. - S. 58-72.

6. Ivakhnenko A. G. Sistemy evristicheskoy samoorganizatsii v tekhnicheskoy kibernetike. - Kiev: "Tekhnika", 1971. - 392 s.

7. Ivakhnenko A. G. Dolgosrochnoe prognozirovanie i upravlenie slozhnymi sistemami. - Kiev: "Tekhnika", 1975. - 311 s.

8. Multivariate adaptive regression splines - Wikipedia [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines

9. Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines". The Annals of Statistics. 19: 1.

10. L1 and L2 Regulatization methods [Електронний ресурс] – Режим доступу до ресурсу: https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

11. Regularization (Mathematics) - Wikipedia [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Regularization_(mathematics)

12. Avoiding overfitting with regulatization [Електронний ресурс] – Режим доступу до ресурсу: https://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/

13. M. Z. Zgurovskiy, A.A. Pavlov, "Prinyatie resheniy v setevykh sistemakh s ogranichennymi resursami: Monografiya", K.: Naukova dumka,–2010.–573 s

14. Zgurovskiĭ M. Z., Pavlov A. A., Misyura Ye. B., Melnikov O. V. Metody operativnogo planirovaniya i prinyatiya reshenii v slozhnykh organizatsi-onno-tekhnologicheskikh sistemakh // Visnik NTUU “KPI”. Informatika, upravlinnya ta obchislyuvalna tekhnika. K.: “VYeK+”, 2010.– No50

15. Pavlov A. A., Kalashnik V. V., Kovalenko D. A. Postroenie bagatovimirnoï polinomialnoï regresiï. Regresiya pri dannykh z povtoryayushchimisya argumentami // Visnik NTUU “KPI”. Seriya «Informatika, upravlinnya ta obchislyuvalna tekhnika». – K.: “VYeK+”, 2015.